

(19)



JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 11250100 A

(43) Date of publication of application: 17 . 09 . 99

(51) Int. Cl

**G06F 17/30**

**G06F 15/18**

**G06F 17/21**

(21) Application number: 10064682

(71) Applicant: NEC CORP

(22) Date of filing: 27 . 02 . 98

(72) Inventor: RI KO  
YAMANISHI KENJI

**(54) HIERARCHICAL DOCUMENT CLASSIFYING DEVICE AND MACHINE-READABLE RECORDING MEDIUM RECORDING PROGRAM**

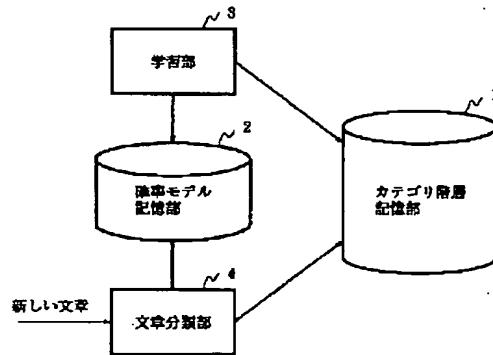
category corresponding to a linear combination model that has the smallest negative logarithmic likelihood.

COPYRIGHT: (C)1999,JPO

(57) Abstract:

**PROBLEM TO BE SOLVED:** To classify a sentence to categorical hierarchies based on the distribution of words (keyword) appearing in the sentence (text and document).

**SOLUTION:** This method grasps document classification as a statistical test problem. A categorical hierarchy storing part 1 stores category hierarchies. A probability model storing part 2 stores linear combination models. A learning part 3 refers to categorical hierarchies stored in the part 1, learns a linear combination model corresponding to each category from a document that is already classified to a category and stores the linear combination model in the part 2. A document classifying part 4 newly inputs a document, refers to each category in category hierarchies stored in the part 1, refers to a linear combination model corresponding to each category from the part 2 for the category, calculates the negative logarithmic likelihood of each linear combination model to the inputted document and classifies the inputted document to a



**THIS PAGE BLANK (USPTO)**

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平11-250100

(43)公開日 平成11年(1999)9月17日

(51)Int.Cl. <sup>6</sup>	識別記号	F I	
G 0 6 F 17/30		G 0 6 F 15/401	3 1 0 D
15/18	5 6 0	15/18	5 6 0 A
17/21		15/20	5 7 0 Z
			5 9 0 Z
		15/40	3 7 0 A

審査請求 有 請求項の数4 FD (全 11 頁)

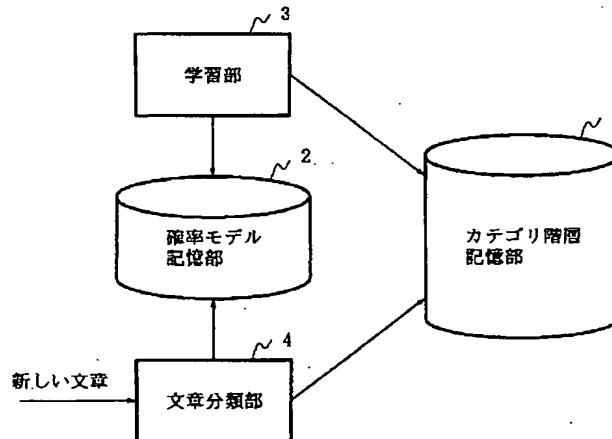
(21)出願番号	特願平10-64682	(71)出願人 000004237 日本電気株式会社 東京都港区芝五丁目7番1号
(22)出願日	平成10年(1998)2月27日	(72)発明者 李 航 東京都港区芝五丁目7番1号 日本電気株式会社内 (72)発明者 山西 健司 東京都港区芝五丁目7番1号 日本電気株式会社内 (74)代理人 弁理士 境 廣巳

(54)【発明の名称】階層型文章分類装置およびプログラムを記録した機械読み取り可能な記録媒体

(57)【要約】

【課題】 文章(テキスト、ドキュメント)に現れる単語(キーワード)の分布を基に文章をカテゴリの階層に分類する。

【解決手段】 文章分類を統計的検定問題として捉える。カテゴリ階層記憶部1ではカテゴリの階層が記憶される。確率モデル記憶部2では、線形結合モデルが記憶される。学習部3は、カテゴリ階層記憶部1に記憶されるカテゴリの階層を参照し、既にカテゴリに分類された文章から各カテゴリの対応する線形結合モデルを学習し、線形結合モデルを確率モデル記憶部2に記憶する。文章分類部4は、新しく文章を入力し、カテゴリ階層記憶部1に記憶されるカテゴリの階層における各カテゴリを参照し、各カテゴリに対して、確率モデル記憶部2から、そのカテゴリに対応する線形結合モデルを参照し、入力文章に対する各線形結合モデルの負対数尤度を計算し、負対数尤度の最も小さい線形結合モデルに対応するカテゴリに入力文章を分類する。



## 【特許請求の範囲】

【請求項1】 ノードが文章の分類されたカテゴリを表現し、リンクがカテゴリの上位下位関係を表現するグラフとして、カテゴリの階層を記憶するカテゴリ階層記憶部、

前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、より下位のカテゴリの単語空間上の確率モデルの重みつき平均を該カテゴリの線形結合モデルとし、各カテゴリの線形結合モデルを記憶する確率モデル記憶部、

前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに分類された文章を基に、各カテゴリの線形結合モデルを、より下位のカテゴリの線形結合モデルから学習し、学習できた各カテゴリの線形結合モデルを前記確率モデル記憶部に記憶する学習部、

新しく文章を入力し、該入力文章を単語のデータ列と見なし、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、前記確率モデル記憶部に記憶される該カテゴリの線形結合モデルの該入力文章に対する負対数尤度を計算し、計算された負対数尤度の最も小さいカテゴリに該入力文章を分類する文章分類部、を備えることを特徴とする階層型文章分類装置。

【請求項2】 ノードが文章の分類されたカテゴリを表現し、リンクがカテゴリの上位下位関係を表現するグラフとして、カテゴリの階層を記憶するカテゴリ階層記憶部、

前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、より下位のカテゴリの、単語空間上の確率モデルの集合を該カテゴリの確率モデルの集合とし、各カテゴリの確率モデルの集合の全ての要素を記憶する確率モデル集合記憶部、

前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに分類された文章を基に、各カテゴリの確率モデルの集合を、より下位のカテゴリの単語空間上の確率モデルの集合から学習し、学習できた各カテゴリの確率モデルの集合のすべての要素を前記確率モデル集合記憶部に記憶する学習部、

新しく文章を入力し、該入力文章を単語のデータ列と見なし、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、前記確率モデル集合記憶部に記憶される該カテゴリの確率モデルの集合に対する該入力文章の確率的複雑度を計算し、計算された確率的複雑度の最も小さいカテゴリに該入力文章を分類する文章分類部、

を備えることを特徴とする階層型文章分類装置。

【請求項3】 コンピュータを、請求項1に記載する、カテゴリ階層記憶部、確率モデル記憶部、学習部、および文章分類部として機能させるプログラムを記録した機械読み取り可能な記録媒体。

【請求項4】 コンピュータを、請求項2に記載する、

2 カテゴリ階層記憶部、確率モデル集合記憶部、学習部、および文章分類部として機能させるプログラムを記録した機械読み取り可能な記録媒体。

## 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】 本発明は、インターネットのホームページの自動分類、電子図書館における文献検索、特許出願情報の検索、電子化された新聞記事の自動分類、マルチメディア情報の自動分類等の情報の分類や検索に関するものである。

## 【0002】

【従来の技術】 情報の分類や検索の分野では、文章分類（ドキュメント分類、テキスト分類ともいう）装置の開発は大きな課題である。ここでいう文章分類とは、予め人間がカテゴリを設け、さらに一部の文章がそれぞれどのカテゴリに属するかを判断し、該当のカテゴリにそれらの文章を分類し、システムに記憶した後、システムは記憶された情報から知識を自動的に獲得し、それ以後、獲得できた知識を基に、新たに入力された文章を自動的に分類することを指す。

【0003】 文章はカテゴリに分類されているので、文章を検索する時、関係するカテゴリにおける文章だけを検索すればよく、検索が効率良く且つ正確になる。

【0004】 従来、幾つかの文章分類装置が提案されている。中でも、Saltonらの提案する文章分類装置が良く知られている（G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, New York: McGraw Hill, 1983）。その文章分類装置は、文章に現れる単語の頻度ベクトルとカテゴリにおける単語の頻度ベクトルとの間のコサイン値を文章とカテゴリ間の距離と見なし、距離の最も小さいカテゴリに文章を分類することを特徴としている。

## 【0005】

【発明が解決しようとする課題】 しかし、従来方式のほとんどは、文章を幾つかの並列のカテゴリに分類するもので、階層構造をなすカテゴリに文章を自動的に分類する装置がなかった。例えば、「政治」のカテゴリがさらに「国会」や「政党」のサブカテゴリに分かれ、文章を「政治」のカテゴリに分類した後、さらにそれを「国会」と「政党」に分類した方が後の検索がさらに高速になる。

【0006】 本発明の目的は、並列のカテゴリに文章を分類するのではなく、階層構造をなすカテゴリに文章を自動分類し得るようにすることにある。

【0007】 また、本発明の別の目的は、信頼性の高い文章の自動分類を実現することにある。

## 【0008】

【課題を解決するための手段】 本発明では、カテゴリを

階層化し、各カテゴリに線形結合モデルと呼ばれる確率モデル、或いは確率モデルの集合を対応させ、新しい文章が入力されると、その文章に対する線形結合モデルの負対数尤度、或いは確率モデル集合の確率的複雑度を計算し、負対数尤度の最も小さい、或いは確率的複雑度の最も小さいカテゴリに新しい文章を分類する。

【0009】つまり、本発明では、文章における単語の分布を基にその文章をカテゴリに分類している。特に、確率的なモデルを用いた統計的検定によって文章を分類することが特徴である。

【0010】具体的には、本発明の第1の階層型文章分類装置は、ノードが文章の分類されたカテゴリを表現し、リンクがカテゴリの上位下位関係を表現するグラフとして、カテゴリの階層を記憶するカテゴリ階層記憶部、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、より下位のカテゴリの単語空間上の確率モデルの重みつき平均を該カテゴリの線形結合モデルとし、各カテゴリの線形結合モデルを記憶する確率モデル記憶部、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに分類された文章を基に、各カテゴリの線形結合モデルを、より下位のカテゴリの線形結合モデルから学習し、学習できた各カテゴリの線形結合モデルを前記確率モデル記憶部に記憶する学習部、新しく文章を入力し、該入力文章を単語のデータ列と見なし、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、前記確率モデル記憶部に記憶される該カテゴリの線形結合モデルの該入力文章に対する負対数尤度を計算し、計算された負対数尤度の最も小さいカテゴリに該入力文章を分類する文章分類部、を備えることを特徴とする。

【0011】このように構成された第1の階層型文章分類装置にあっては、学習部が、カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに例えれば事前に人手によって分類された文章を基に、各カテゴリの線形結合モデルを、より下位のカテゴリの線形結合モデルから学習し、学習できた各カテゴリの線形結合モデルを確率モデル記憶部に記憶し、その後、自動分類対象となる文章が入力されると、文章分類部が、その文章を入力し、この入力文章を単語のデータ列と見なし、カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、確率モデル記憶部に記憶される該カテゴリの線形結合モデルの該入力文章に対する負対数尤度を計算し、計算された負対数尤度の最も小さいカテゴリに該入力文章を分類する。

【0012】また、本発明の第2の階層型文章分類装置は、ノードが文章の分類されたカテゴリを表現し、リンクがカテゴリの上位下位関係を表現するグラフとして、カテゴリの階層を記憶するカテゴリ階層記憶部、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、より下位のカテゴリの、単語空間上の確

率モデルの集合を該カテゴリの確率モデルの集合とし、各カテゴリの確率モデルの集合の全ての要素を記憶する確率モデル集合記憶部、前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに分類された文章を基に、各カテゴリの確率モデルの集合を、より下位のカテゴリの単語空間上の確率モデルの集合から学習し、学習できた各カテゴリの確率モデルの集合のすべての要素を前記確率モデル集合記憶部に記憶する学習部、新しく文章を入力し、該入力文章を単語のデータ列と見なし、

10 前記カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、前記確率モデル集合記憶部に記憶される該カテゴリの確率モデルの集合に対する該入力文章の確率的複雑度を計算し、計算された確率的複雑度の最も小さいカテゴリに該入力文章を分類する文章分類部、を備える。

【0013】このように構成された第2の階層型文章分類装置にあっては、学習部が、カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに例えれば事前に人手によって分類された文章を基に、各カテゴリの確率モ

20 デルの集合を、より下位のカテゴリの単語空間上の確率モデルの集合から学習し、学習できた各カテゴリの確率モデルの集合のすべての要素を確率モデル集合記憶部に記憶し、その後、自動分類対象となる文章が入力されると、文章分類部が、その文章を入力し、この入力文章を単語のデータ列と見なし、カテゴリ階層記憶部に記憶されるカテゴリの階層の各カテゴリに対して、確率モデル集合記憶部に記憶される該カテゴリの確率モデルの集合に対する該入力文章の確率的複雑度を計算し、計算された確率的複雑度の最も小さいカテゴリに該入力文章を分類する。

【0014】

【発明の実施の形態】次に本発明の実施の形態の例について図面を参照して詳細に説明する。

【0015】図1を参照すると、本発明の第1の実施例は、カテゴリ階層記憶部1、確率モデル記憶部2、学習部3、および文章分類部4から構成される。

【0016】カテゴリ階層記憶部1ではカテゴリの階層が記憶される。カテゴリの階層構造はグラフとして表される。グラフでは、ノードがカテゴリを表現し、リンク

40 がカテゴリの上位下位関係を表現する。また、カテゴリには既に分類された文章が入っている。図2にカテゴリ階層の例を示す。ここでは、カテゴリの階層が木構造となっているが、一般的にはカテゴリの階層がもっと複雑なグラフ構造になる。

【0017】確率モデル記憶部2では、カテゴリの階層における一つのカテゴリに対して一つの確率モデルを対応させて記憶する。本実施例では、この確率モデルが線形結合モデルの形をとることを特徴とする。ある確率の線形結合モデルは、それより下位のカテゴリの確率モデルの重みつき平均として定義される。以下に線形結合モ

モデルの例を示す。

【0018】線形結合モデルの例；カテゴリの木構造では、ノードがカテゴリを表す。ノードcのカテゴリの線形結合モデルはその子ノードのカテゴリの線形結合モデル

$$P(W | c) = \sum_{i=1}^n P(W | c_i) \times P(c_i | c) + P(W | c') \times P(c' | c)$$

数1において、確率変数Wは単語の集合W = {w1, w2, …, ws} の値をとる。P(W | c1), P(W | c2), …, P(W | cn) はcの子ノードc1, c2, …, cnのカテゴリの線形結合モデルである。P(W | c') はノードc自身に属する確率モデルである。つまり、P(W | c') はcの表すカテゴリに属し、c1, …, cnの表すカテゴリに属さない確率モデルである。P(c' | c), P(c1 | c), …, P(cn | c) はc', c1, …, cnの事前確率である。

【0019】学習部3は、カテゴリ階層記憶部1に記憶されるカテゴリの階層を参照し、既にカテゴリに分類された文章から各カテゴリの線形結合モデルを学習し、学習できた線形結合モデルを確率モデル記憶部2に記憶する。

【0020】文章分類部4は、新しく文章を入力し、該文章を単語のデータ列と見なし、カテゴリ階層記憶部1に記憶されるカテゴリの階層における各カテゴリを参照

ノードcを入力とする。最初は、木構造のルートノードが入力される。

if

ノードcは葉ノードである。

then

ノードcのカテゴリに分類された30章から、cの線形結合モデルを学習し、戻る。

else

ノードcの子ノードci (i = 1, 2, …, n) の線形結合モデルを参照する。

if

ノードciの線形結合モデルはまだ学習できていない。

then

ノードciに対して、再帰的に本アルゴリズムを適用する。

else

ノードciの線形結合モデルとc自身の確率モデルからノードcの線形結合モデルを学習し、戻る。

【0024】文章分類部4は文章の統計的仮説検定によって文章を分類する。次に、文章分類部4のアルゴリズム

dは入力された文章であるとする。ノードcと文章dを入力とする。最初は、木構造のルートノードが入力される。

if

ノードcは葉ノードである。

then

文章dはノードcのカテゴリに属するとし、終了する。

else

し、各カテゴリに対して、確率モデル記憶部2から、そのカテゴリの対応する線形結合モデルを参照し、該文章に対する各線形結合モデルの負対数尤度を計算し、負対数尤度のもっとも小さい線形結合モデルに対応するカテゴリに該文章を分類する。

【0021】学習部3は、幾つかの方法で線形結合モデルを学習することができる。例えば、その下位カテゴリの線形結合モデルをヒストグラムとして推定することができる。また、重み係数をEMアルゴリズムと呼ばれるアルゴリズムによって学習することができる。

【0022】ここでは、学習部3の学習アルゴリズムの一例を示す。階層を表すグラフは木構造をもつとする。学習部3は、木構造となるカテゴリの階層を参照し、ボトムアップ的にカテゴリの線形結合モデルを学習する。その学習アルゴリズムは以下の通りであり、そのフローチャートを図3に示す。

【0023】

ノードcを入力とする。最初は、木構造のルートノードが入力される。

if

ノードcは葉ノードである。

then

ノードcのカテゴリに分類された30章から、cの線形結合モデルを学習し、戻る。

else

ノードcの子ノードci (i = 1, 2, …, n) の線形結合モデルを参照する。

if

ノードciの線形結合モデルはまだ学習できていない。

then

ノードciに対して、再帰的に本アルゴリズムを適用する。

else

ノードciの線形結合モデルとc自身の確率モデルからノードcの線形結合モデルを学習し、戻る。

【0025】

ムの一例を示し、そのフローチャートを図4に示す。

7

8

文章  $d$  に対するノード  $c$  の線形結合モデルの負対数尤度  $L(d | c)$  を計算する。ノード  $c$  の子ノード  $c_i$  ( $i = 1, 2, \dots, n$ ) の負対数尤度  $L(d | c_i)$  をも計算する。計算できた  $L(d | c)$  と  $L(d | c_i)$  の最小値を求める。

if

子ノードの中の  $c_i$  の負対数尤度が最小である。

then

ノード  $c_i$  に対して本アルゴリズムを再帰的に適用する。

else

文章  $d$  はノード  $c$  のカテゴリに属するとし、終了する。

【0026】次に、学習部3による線形結合モデルを学習する方法と、文章分類部4による負対数尤度の計算方法を、さらに具体的な例を通じて説明する。カテゴリの階層は図5に示すものとする。図5中、 $c_1, c_2, c_3$  はカテゴリであり、 $d_1, d_2, d_3$  は既に分類された文章である。また、図6に各文章  $d_1, d_2, d_3$  における単語  $w_1, w_2, w_3$  の出現頻度を示す。単語  $w_1, w_2, w_3$  は予め定められたキーワードである。

【0027】○線形結合モデルの学習の例

$c_2$  と  $c_3$  は葉ノードであるので、それらのノードのカテゴリの線形結合モデルは文章における単語のヒストグラムとして、図7(a)のように学習される。

【0028】 $c_1$  に分類された文章  $d_2$  から、 $c_1$  自身に属する確率モデルを単語のヒストグラムとして学習する。これを  $P(w | c_1')$  と表す。つまり、それは  $c$

$$P(w_1 | c_1') = P(w_1 | c_1') \times P(c_1' | c_1) + P(w_2 | c_1') \times P(c_2 | c_1) + P(w_3 | c_1') \times P(c_3 | c_1)$$

$$= \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} = \frac{4}{9}$$

1のカテゴリに属し、 $c_2, c_3$  のカテゴリに属さない確率モデルであり、図7(b)のように学習される。

【0029】一方、各モデルの事前分布を以下のように学習する。

【数2】

$$P(c_i | c) = \frac{f(c_i)}{N}$$

【0030】ここで、 $f(c_i)$  はノード  $c_i$  とその支配するノードの属する文章数で、 $N$  は全文章数である。

20 よって、各モデルの事前分布は図7(c)のように学習される。

【0031】さらに、線形結合モデルの定義に従って、ノード  $c_1$  における線形結合モデルを以下のように学習することができる。

【数3】

$$P(w_2 | c_1) = P(w_2 | c_1') \times P(c_1' | c_1) + P(w_2 | c_2) \times P(c_2 | c_1) + P(w_2 | c_3) \times P(c_3 | c_1)$$

$$= \frac{1}{4} \times \frac{1}{3} + \frac{1}{6} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} = \frac{11}{36}$$

$$P(w_3 | c_1) = P(w_3 | c_1') \times P(c_1' | c_1) + P(w_3 | c_2) \times P(c_2 | c_1) + P(w_3 | c_3) \times P(c_3 | c_1)$$

$$= \frac{1}{4} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{6} \times \frac{1}{3} = \frac{1}{4}$$

【0032】即ち、ノード  $c_1$  における線形結合モデルは図7(d)に示すようになる。

【0033】○負対数尤度の計算の例

新しい文章  $d$  における単語の分布は図8に示すものとする。つまり、文章分類部4は入力文章中から単語  $w_1$  を2個、単語  $w_2$  を1個、単語  $w_3$  を1個検出したとする。 $d$  に対する  $c_1$  の負対数尤度を以下のように計算する。対数の底は2であるとする。

【数4】

$$L(d | c_1) = -\log \frac{4}{9} - \log \frac{11}{36} - \log \frac{1}{4} = 6.05$$

【0034】同様に、 $c_2, c_3$  の負対数尤度を計算する。

$$L(d | c_2) = 2 \cdot -\log \frac{1}{2} - \log \frac{1}{6} - \log \frac{1}{3}$$

$$= 6.17$$

【数6】

$$L(d | c_3) = 2 \cdot -\log \frac{1}{3} - \log \frac{1}{2} - \log \frac{1}{6}$$

$$= 6.75$$

【0035】尤度  $L(d | c_1)$  がもっとも小さいので、dはc<sub>1</sub>に分類される。

【0036】図9を参照すると、本発明の第2の実施例は、カテゴリ階層記憶部1、確率モデル集合記憶部5、学習部6、および文章分類部7から構成される。

【0037】カテゴリ階層記憶部1ではカテゴリの階層が記憶される。カテゴリの階層では、ノードがカテゴリを表し、リンクが上位下位関係を表す。カテゴリ階層の例として前述した図2がある。

【0038】確率モデル集合記憶部5では、確率モデルの集合が記憶される。カテゴリの階層における各カテゴリに対して一つの確率モデルの集合が定義され、記憶される。以下に確率モデルの集合の例を示す。

#### 【0039】○確率モデル集合の例

ノードcの確率モデルの集合が確率モデル  $P(w | c')$ ,  $P(w | c_1)$ , ...,  $P(w | c_n)$  を含むとする。 $P(w | c_1)$ , ...,  $P(w | c_n)$  はcの子ノード  $c_1, \dots, c_n$  の確率モデルの集合のもつ確率モデル（確率分布）である。 $P(w | c')$  はノードc自身に属する確率モデルである。つまり、それは、cのカテゴリに属し、 $c_1, \dots, c_n$  のカテゴリに属さない確率モデルである。また、各確率モデルの事前確率  $P(c' | c)$ ,  $P(c_1 | c)$ , ...,  $P(c_n | c)$  が存在するとする。確率モデル  $P(w | c')$ ,  $P(w | c_1)$ , ...,  $P(w | c_n)$  は、例えば、ヒストグラムの形で表現される。

【0040】各カテゴリの確率モデルの集合は、それ自身に属する文章による単語空間上の確率モデルと、その下位のカテゴリに属する文章による単語空間上の確率モデルからなる。

ノードcを入力とする。最初は、木構造のルートノードが入力される。  
if

ノードcは葉ノードである。

then

ノードcのカテゴリに分類された文章から、cの確率モデル集合の全ての要素を学習し、戻る。

else

ノードcの子ノードc<sub>i</sub> ( $i = 1, 2, \dots, n$ ) の確率モデル集合を参照する。

if

ノードc<sub>i</sub>の確率モデル集合はまだ学習できていない。  
then

【0041】学習部6は、カテゴリ階層記憶部1に記憶されるカテゴリの階層を参照し、既にカテゴリに分類された文章から各カテゴリの対応するモデル集合を学習し、学習できた確率モデルの集合を確率モデル集合記憶部5に記憶する。

【0042】文章分類部7は、新しく文章を入力し、該文章を単語のデータ列と見なし、カテゴリ階層記憶部1に記憶されるカテゴリにおける各カテゴリを参照し、各カテゴリに対して、確率モデル集合記憶部5から、そのカテゴリの対応する確率モデル集合を参照し、該文章の各参照された確率モデル集合に対する確率的複雑度を計算し、確率的複雑度のもっとも小さい確率モデル集合に対応するカテゴリに該文章を分類する。

【0043】確率的複雑度とは、確率モデルの集合を用いてデータを記述する際の最小記述長を表す量で、リッサン (Rissanen) によって提唱されたものである (Jorma Rissanen, Stochastic Complexity in Statistical Inquiry, World Scientific Publishing Co., Singapore, 1989)。本実施例では、確率的複雑度を、確率モデル集合における確率モデルのデータに対する尤度の重み付き平均の負対数として計算する。

【0044】次に、学習部6の学習アルゴリズムの一例を示す。階層を表すグラフが木構造をもつとする。学習部6は、木構造となるカテゴリの階層を参照し、ボトムアップ的にカテゴリの確率モデル集合を学習する。そのアルゴリズムは以下の通りであり、そのフローチャートを図10に示す。

#### 【0045】

11

12

ノード  $c_i$  に対して、再帰的に本アルゴリズムを適用する。

**e l s e**

ノード  $c_i$  の確率モデル集合と  $c$  に分類された文章の確率モデルからノード  $c$  の確率モデル集合を学習し、戻る。

【0046】文章分類部7は統計的仮説検定によって文章を分類する。次に、文章分類部7のアルゴリズムの一

例を示す。図11はそのフローチャートである。

【0047】

$d$  は入力された文章であるとする。ノード  $c$  と文章  $d$  を入力とする。最初は、木構造のルートノードが入力される。

**i f**

ノード  $c$  は葉ノードである。

**t h e n**

文章  $d$  はノード  $c$  のカテゴリに属するとし、終了する。

**e l s e**

ノード  $c$  における文章  $d$  の確率的複雑度  $SC(d | c)$  を計算する。ノード  $c$  の子ノード  $c_i$  ( $i = 1, 2, \dots, n$ ) における確率的複雑度  $SC(d | c_i)$  をも計算する。計算できた  $SC(d | c)$  と  $SC(d | c_i)$  の中の最小値を求める。

**i f**

ノードの中の  $c_i$  の確率的複雑度が最小である。

**t h e n**

ノード  $c_i$  に対して本アルゴリズムを再帰的に適用する。

**e l s e**

文章  $d$  はノード  $c$  のカテゴリに属するとし、終了する。

【0048】次に確率的複雑度の計算例を示す。

【0049】カテゴリの階層は図5に示すものとする。

また、文章における単語（キーワード）の出現頻度は図6に示すものであるとする。

【0050】ノード  $c_2, c_3$  が葉ノードであるので、それぞれのもつ確率モデルの集合は一つの確率モデルを含む。さらに、それらの確率モデルがヒストグラムとして、図12(a)のように学習される。

【0051】ノード  $c_1$  自身に属する確率モデルもヒストグラムとして、図12(b)のように学習される。

$$SC(d | c_1) = -\log \left( \frac{1}{3} \cdot \left( \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{6} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{6} \right) \right)$$

【0055】 $d$  の  $c_2, c_3$  に対する確率的複雑度を以下のように計算する。

【数8】

$$SC(d | c_2) = -\log \left( \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{6} \right) \\ = 6.75$$

【数9】

$$SC(d | c_3) = -\log \left( \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{6} \cdot \frac{1}{3} \right) \\ = 6.17$$

【0056】 $SC(d | c_3)$  がもっとも小さいので、 $d$  は  $c_3$  に分類される。

【0057】図13は本発明の階層型文章分類装置の第3の実施例のブロック図である。この例の階層型文章分類装置は、CPU101、主記憶102および補助記憶

【0052】従って、 $c_1$  の確率モデル集合は確率モデル  $P(W | c_1), P(W | c_2), P(W | c_3)$  を含むことになる。それらの確率モデルの事前確率  $P(d | i | c)$  が一様分布であるとする。

【0053】新しい文章  $d$  における単語の出現頻度は図12(c)に示すものであるとする。

【0054】 $d$  の  $c_1$  に対する確率的複雑度を以下のように計算する。対数の底は2であるとする。

【数7】

103を含むコンピュータ104と、このコンピュータ104に接続された表示装置105、入力装置106およびファイル107を含む入出力装置108と、階層型文章分類プログラムを記録する記録媒体109とから構成される。記録媒体109はCD-ROM、半導体メモリ等の機械読み取り可能な記録媒体であり、ここに記録された階層型文章分類プログラムは、コンピュータ104に読み取られ、コンピュータ104の動作を制御することにより、コンピュータ104上に、図1に示したカテゴリ階層記憶部1、確率モデル記憶部2、学習部3および文章分類部4、または図9に示したカテゴリ階層記憶部1、確率モデル集合記憶部5、学習部6および文章分類部7を実現する。

【0058】

【発明の効果】以上説明したように、本発明によれば、

階層構造をなすカテゴリに文章を自動分類することができ、かつ尤度比検定の理論に基づいた統計的信頼性の高い文章分類ができる。

【図面の簡単な説明】

【図1】本発明の階層型文章分類装置の第1の実施例のブロック図である。

【図2】カテゴリ階層の例を示す図である。

【図3】本発明の階層型文章分類装置の第1の実施例における学習アルゴリズムの一例を示すフローチャートである。

【図4】本発明の階層型文章分類装置の第1の実施例における文章分類のアルゴリズムの一例を示すフローチャートである。

【図5】カテゴリ階層の例を示す図である。

【図6】文章における単語分布の例を示す図である。

【図7】線形結合モデルの学習例の説明図である。

【図8】負対数尤度の計算例の説明図である。

【図9】本発明の階層型文章分類装置の第2の実施例の

ブロック図である。

【図10】本発明の階層型文章分類装置の第2の実施例における学習アルゴリズムの一例を示すフローチャートである。

【図11】本発明の階層型文章分類装置の第2の実施例における文章分類のアルゴリズムの一例を示すフローチャートである。

【図12】確率的複雑度の計算例の説明図である。

【図13】本発明の階層型文章分類装置の第3の実施例のブロック図である。

【符号の説明】

1 カテゴリ階層記憶部

2 確率モデル記憶部

3 学習部

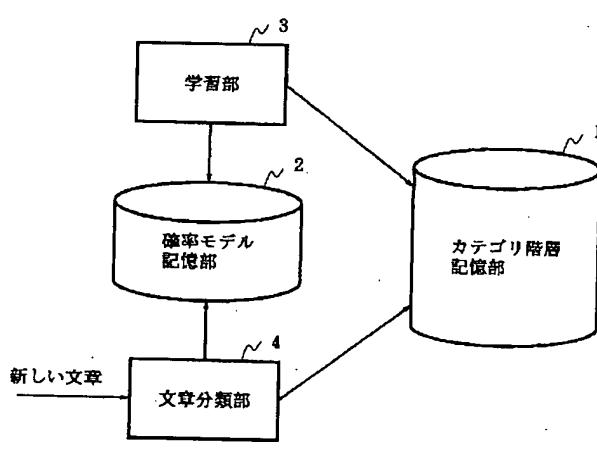
4 文章分類部

5 確率モデル集合記憶部

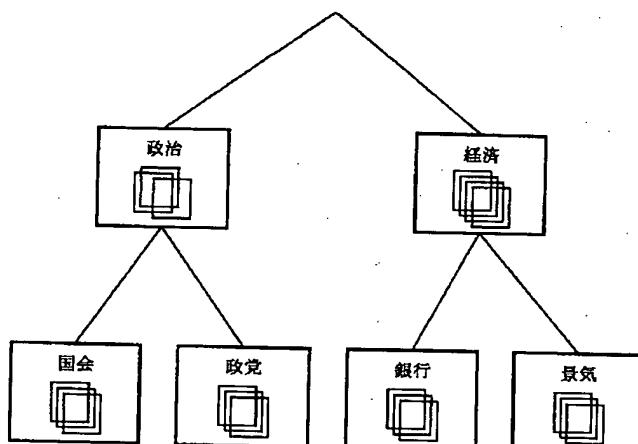
6 学習部

7 文章分類部

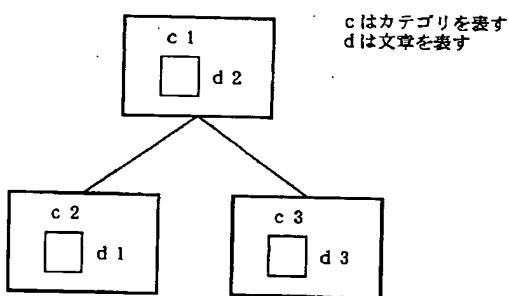
【図1】



【図2】



【図5】



【図8】

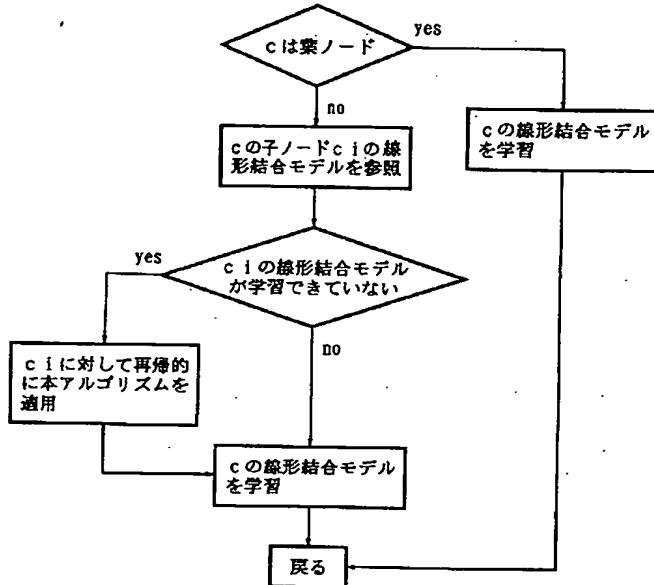
	w1	w2	w3
f(w)	2	1	1

: カテゴリ

: 分類された文章

【図3】

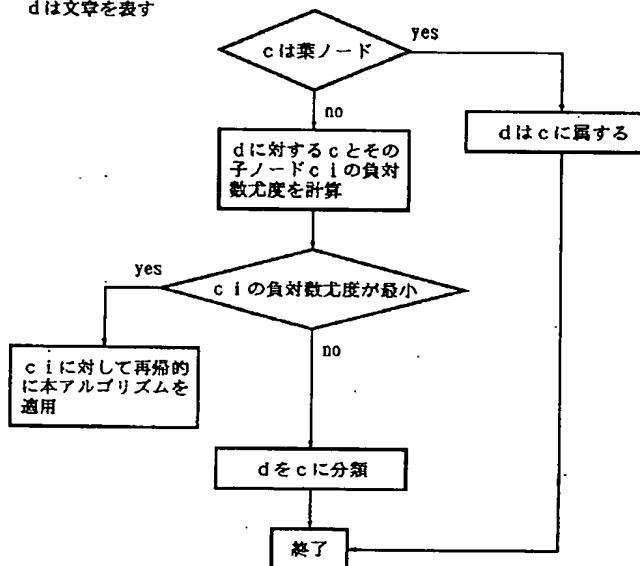
cはノードを表す



【図6】

f(W D)	w1	w2	w3
d1	2	0	1
d2	1	0	0
d3	1	2	0

【図4】

cはノードを表す  
dは文章を表す

【図7】

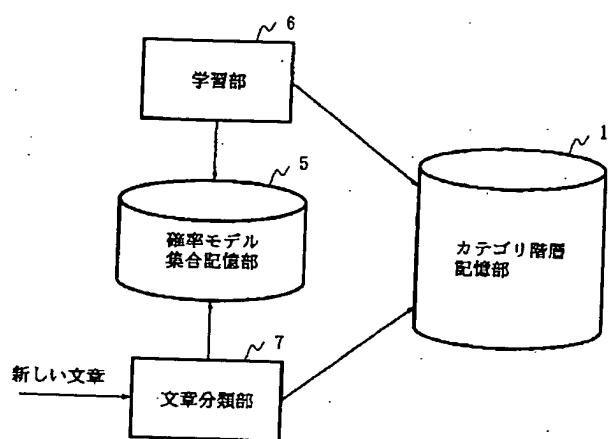
	w1	w2	w3
P(W c2)	1/2	1/6	1/3
P(W c3)	1/3	1/2	1/6

	w1	w2	w3
P(W c1')	1/2	1/4	1/4

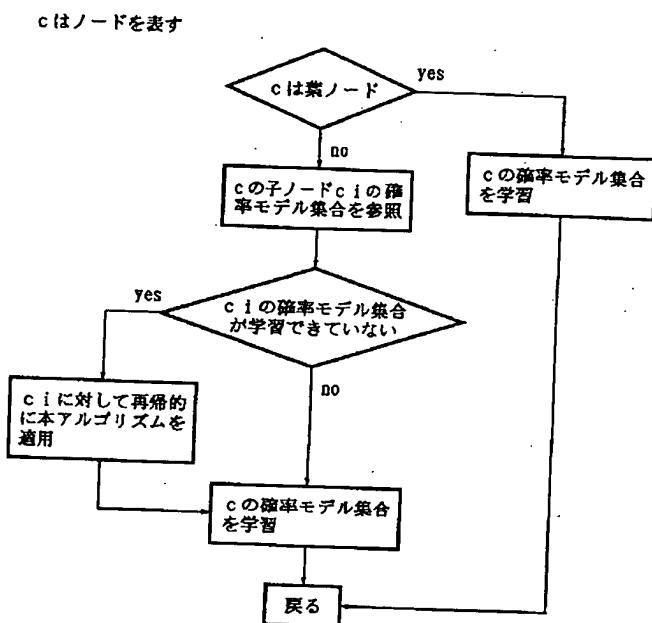
$$\begin{aligned}
 P(c_2 | c_1) &= 1/3 \\
 P(c_3 | c_1) &= 1/3 \\
 P(c_1' | c_1) &= 1/3
 \end{aligned}$$

	w1	w2	w3
P(W c1)	4/9	11/36	1/4

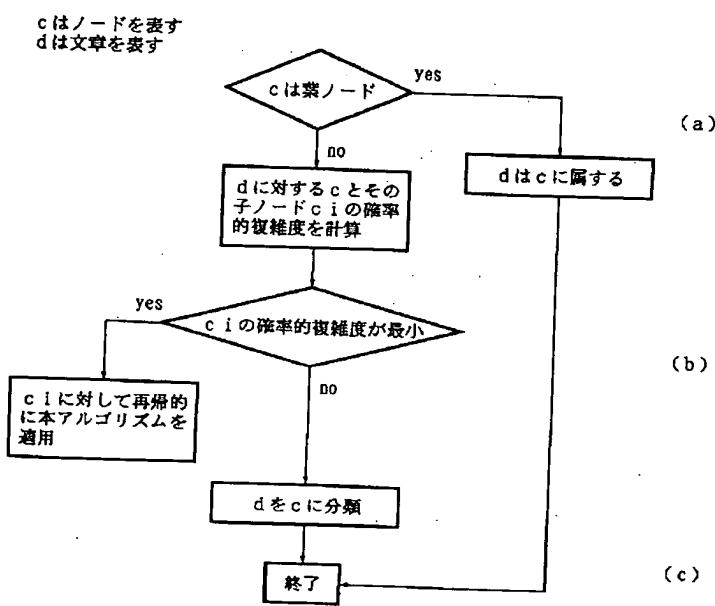
【図9】



【図10】



【図11】



【図12】

	w <sub>1</sub>	w <sub>2</sub>	w <sub>3</sub>
P(W c <sub>2</sub> )	1/2	1/6	1/3
P(W c <sub>3</sub> )	1/3	1/2	1/6

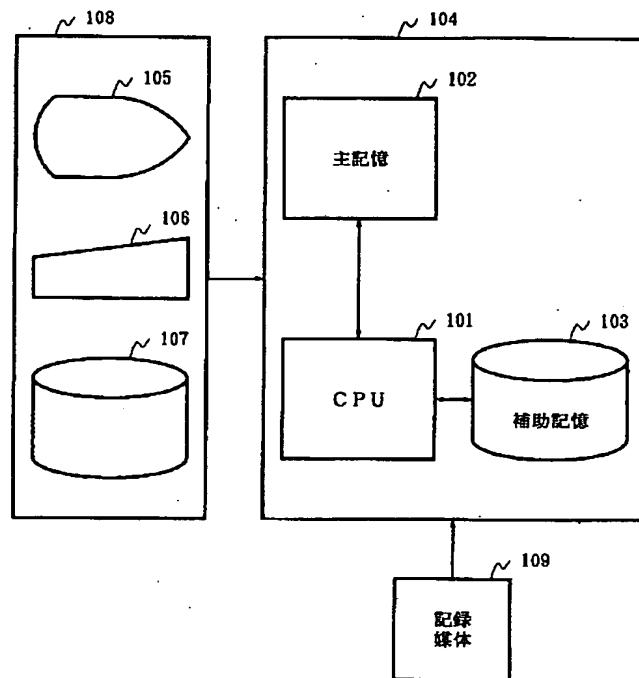
  

	w <sub>1</sub>	w <sub>2</sub>	w <sub>3</sub>
P(W c <sub>3</sub> )	1/2	1/4	1/4

	w <sub>1</sub>	w <sub>2</sub>	w <sub>3</sub>
f(W d <sub>3</sub> )	2	1	1

【図13】



THIS PAGE BLANK (USPTO)